OSDN: an Open Science Data Network for Interdisciplinary Research (demonstration)

Vincent-Nam Dang, Nathalie Aussenac-Gilles, Imen Megdiche, Franck Ravat

OSDN: an Open Science Data Network for Interdisciplinary Research

Vincent-Nam Dang[1,2], Nathalie Aussenac-Gilles[1][0000−0003−3653−3223], Imen Megdiche[1,3][0000−0002−1331−8662], and Franck Ravat[1,2][0000−0003−4820−841X]

[1] IRIT, CNRS (UMR 5505), Université de Toulouse, France
[2] Université Toulouse Capitole, France
[3] INU Champollion, ISIS Castres, Université de Toulouse, France

Abstract. Open Science is a movement in scientific research aimed at opening up data and knowledge creation processes to enable new collaborations. Feedback on Open Science deployment describes an absence of inter and intra-community coordination solution. To overcome the lack of data management coordination in Open Science, we propose the Open Science Data Network (OSDN). The OSDN allows an interconnection of Open Science data management platforms and implements an interdisciplinary information exchange. The OSDN is evaluated on a realistic scenario based on existing platforms and demonstrating its benefits to facilitate data findability and access from one data portal to another, even in a cross-disciplinary context.

Keywords: Interdisciplinary Information Exchange, Open Science, Distributed and Decentralized Architecture, Interoperability

1 Introduction

Open Science is a movement in scientific research aimed at opening up data and knowledge creation processes to enable new collaborations between researchers. The first feedback on issues limiting the generalization of Open Science describes 4 major problems [8]: (i) absence of an inter- and intra-community coordination solution; (ii) lack of access to technical support and lack of resources (time, human resources, training) to help researcher on opening their data [9,7,6]; (iii) an high diversity of needs from researchers; (iv) no security, confidentiality and control on data from centralized solutions [9,7,4,6]. There is a trend towards standardization and centralization of proposed solutions, particularly in terms of metadata models. Schema.org is the model favored by certain communities, notably Google Data Search [4]. This standardization approach can also be found in data management platform projects, based on pivotal models such as Agora [10] or OCEAN protocol [5]. However, unification to one metadata model cannot meet all needs or be applied across all existing communities [2]. To implement an information exchange, a first approach relies on centralized data management at institutional level. Such platforms are numerous (more than 3000 on re3data.org).

[4] https://datasetsearch.research.google.com/ [5] oceanprotocol.com/

Each of them targets different communities – in relation to domains[6] [7], targeted institutions or researchers with knowledge of them. But centralization does not answer the problematic of interdisciplinary information exchange and create a high diversity of solutions. To implement information exchange in Open Science, it is necessary to take this diversity into account. This paper focuses on describing a solution to answer the lack of interdisciplinary information exchange in Open Science with Open Science Data Network (OSDN), based on decentralization and integration of existing data management platforms to let researcher use platforms they already use.

2 Open Science Data Network – OSDN

We propose OSDN, a decentralized federated network of data management platforms. This network is defined by the interconnection of exisiting platform with the integration of a module (see Fig 1). This module is based on a RESTful API in a Docker container, allowing a simplified and automated deployment. The integration of the module in the platform is done through the adding of a single function that allows the module to make requests on the platforms. This module is based on a registry shared by all platform in the OSDN, containing platforms informations, metadata models used by platforms and matches between models. The exchange protocol implemented include a propagation query mecanism based on a standardized format of messages and a recursive broadcast to all neighboor until every platform received the query. The OSDN is based on our formal model of interoperability, that is divided in 2 type of interoperability with technical interoperability and semantic interoperability [3]. Technical interoperability is implemented with standardization of exchange protocol and semantic interoperability is based gateway implementation on metadata models with matches. Theses matches can be done manually, like with crosswalk developed by RDA[8], or automatically with automatic matching algorithm, like the ones we observe in OAEI[9]. A transitive closure is applied to matches to reduce the cost of matching models. As the information exchange is a critical point in knowledge creation process, robustness and durability of the network have to be included in conception. We use a scale-free network topology, following a power-law with a parameter $2 < \gamma < 3$ [1]. This topology allow to minimize the impact of attacks or random deletion when the minimal degree is 5 and allow to have a reduced propagation time of information [1]. We decided to set a minimal mandatory degree of 2, to reduce cost adoption. To follow this topology, we added an inscription protocol, that provide the first platform to connect to, that minimize the Kullback-Leibner divergence, and letting the second to be freely selected. The inscription defined which platform the request will be send first and the model that have to be matched with which the local model have to be matched.

[6] www.rcsb.org [7] beacon-network.org [8] `www.rd-alliance.org/group/research-metadata-schemas-wg/outcomes/collection-crosswalks-fifteen-research-data-schemas` [9] https://oaei.ontologymatching.org/2023/

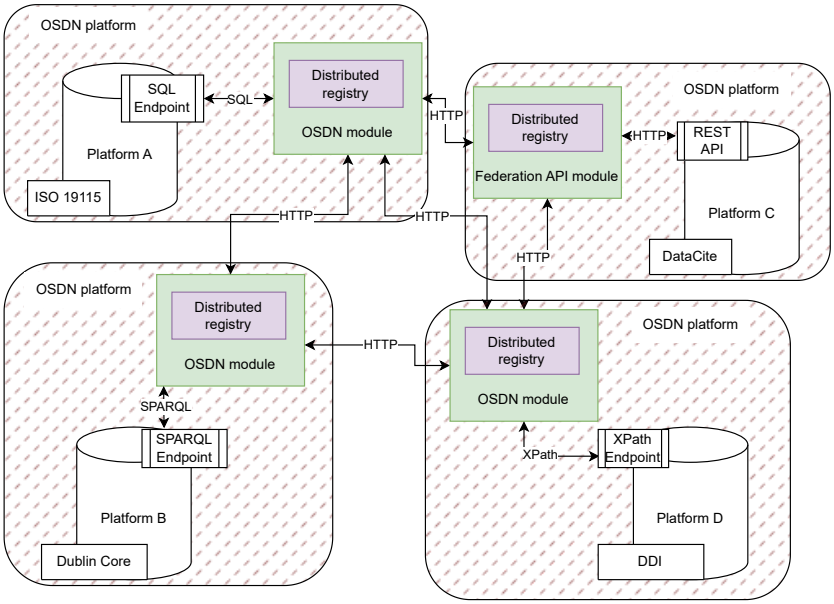

Fig. 1: OSDN network - example with 4 data platforms

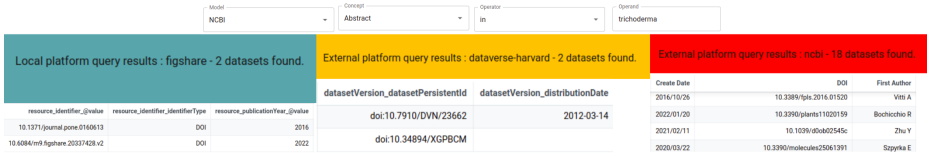

Fig. 2: Results request on description containing "trichoderma"

3 Results

We implemented a POC of OSDN with a WEB GUI to query datasets in OSDN[10]. We integrated 11 existings platforms from differents community and domains, to represent the diveristy of Open Science (AERIS, ForM@Ter and Theia from DataTerra, Bacterial and Viral Bioinformatics Resource Center, data platform of European Union, RCSB Protein Data Bank, Figshare, Harvard Dataverse, The Humanitarian Data Exchange, American National Library of Medecine, Opendatasoft). Dr. Pressecq[0000−0003−0067−7903] researcher on agronomy evaluated our platform. In his research field, he has problems accessing relevant data to create decision support system for farmer to use biocontrol tools [5]. He query OSDN with the same he used for his research (on trichoderma harzianum T-22 strain, see Fig. 2). OSDN showed that it would allow in his research project to gain 80% time on dataset creation on ≈ 10% of his

[10] `https://github.com/vincentnam/OSDN_WEB_GUI_experiment`

datasets, allowed an increasing of 7% in data volume and allowed him to define several new interdisciplinary future works based on datasets found in OSDN.

4 Conclusion

OSDN is a decentralized data management interconnection network allowing to implement a interdisciplinary and intercommunity information research. OSDN is based on a shared registry and community efforts of whole Open Science community. We evaluated our POC on a real research project and we identified the contributions of OSDN, saving time in building datasets, enirching data, allowing datasets reuses and providing new interdisciplinary research perspectives. For future work, we plan to (i) enhance semantic interoperability, automatic matching to scale up with Open Science characteristics and add more semantics relations between concepts; (ii) optimize the implementation of the registry which still fully replicated across all platforms (iii) optimize and reduce cost of query propagation.

Aknowledgment

We thank Dr. T. Pressecq from INRAE and APREL for his participation in our experiment and his feedback.

References

1. Barabási, A.L., Pósfai, M.: Network science. Cambridge University Press, Cambridge (2016), http://barabasi.com/networksciencebook/
2. Chan, L.M., Zeng, M.L.: Metadata interoperability and standardization–a study of methodology part i. D-Lib magazine **12**(6), 1082–9873 (2006)
3. Dang, V.N., Aussenac-Gilles, N., Megdiche, I., Ravat, F.: Interoperability of open science metadata: What about the reality? In: RCIS2023. pp. 467–482. Springer Nature, Cham (2023)
4. Kathawalla, U.K., Silverstein, P., Syed, M.: Easing into open science: A guide for graduate students and their advisors. Collabra: Psychology **7**(1), 18684 (2021)
5. Pressecq, T., et al.: Developing decision support systems based on protection efficacy to promote the use of microbial biocontrol agents in the field (2020)
6. Rainey, L., et al.: Fair data sharing: An international perspective on why medical researchers are lagging behind. BDS **10**(1), 20539517231171052 (2023)
7. Sadeh, Y., et al.: Opportunities for improving data sharing and fair data practices to advance global mental health. Cambridge Prisms: GMH **10**, e14 (2023)
8. Science, D., Hahnel, M., Smith, G., henning schoenenberger, Scaplehorn, N., Day, L.: The State of Open Data 2023. Digital Science (11 2023). https://doi.org/10.6084/m9.figshare.24428194.v1
9. Top, J., Janssen, S., Boogaard, H., Knapen, R., Şimşek-Şenel, G.: Cultivating fair principles for agri-food data. Computers and Electronics in Agriculture **196**, 106909 (2022)
10. Traub, J., Kaoudi, Z., Quiané-Ruiz, J.A., Markl, V.: Agora: Bringing together datasets, algorithms, models and more in a unified ecosystem [vision]. ACM SIGMOD Record **49**(4), 6–11 (2021)